Estimation

The ASTA team

Contents

1 Point and interval estimates

1.1 Point and interval estimates

- Suppose we study a population and we are interested in certain parameters of the population distribution, e.g. the mean μ and the standard deviation σ.
- Based on a sample we can make a **point estimate** of the parameter. We have already seen the following examples:
  - $\bar{x}$ is a point estimate of μ
  - s is a point estimate of σ

- We often supplement the point estimate with an **interval estimate** (also called a **confidence interval**). This is an interval around the point estimate, in which we are confident (to a certain degree) that the population parameter is located.

1.2 Point estimators: Bias

- If we want to estimate the population mean μ we have several possibilities e.g.
  - the sample mean $\bar{X}$
  - the average X_T of the sample upper and lower quartiles
- Advantage of X_T: Very large/small observations have little effect, i.e. it has practically no effect if there are a few errors in the data set.
- Disadvantage of X_T: If the distribution of the population is skewed, i.e. asymmetrical, then X_T is **biased**, i.e. $E(X_T) \neq \mu$. This means that in the long run this estimator systematically over or under estimates the value of μ.
- Generally we prefer that an estimator is **unbiased**, i.e. its expected value equals the true parameter value.
- Recall that for a sample from a population with mean μ, the sample mean $\bar{X}$ also has mean μ. That is, $\bar{X}$ is an unbiased estimate of the population mean μ.

1.3 Point estimators: Consistency

- From previous lectures we know that the standard error of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$, so
  - the standard error decreases when the sample size increases.
- In general an estimator with this property is called **consistent**.
- X_T is also a consistent estimator, but has a variance that is greater than $\bar{X}$.

1.4 Point estimators: Efficiency

- Since the variance of X_T is greater than the variance of $\bar{X}$, we prefer $\bar{X}$.
- In general, we prefer the estimator with the smallest possible variance. This estimator is said to be **efficient**.
- $\bar{X}$ is an efficient estimator.

1.5 Notation

- The symbol $\hat{}$ above a parameter denotes a (point) estimate of the parameter. We have looked at an
  - estimate of the population mean μ, namely $\hat{\mu} = \bar{x}$.
  - estimate of the population standard deviation σ, namely $\hat{\sigma} = s$
  - estimate of the population proportion p, namely the sample proportion $\hat{p}$.

2 Confidence intervals

2.1 Confidence Interval

- A **confidence interval** for a parameter is constructed as an interval, where we expect the population parameter to be.
- The probability that this construction yields an interval which includes the population parameter is called the **confidence level**.
  - We write the confidence level as $100(1-\alpha)\%$.
  - The confidence level is typically chosen to be 95%.
  - α is called the **error probability**.
  - For a 95% confidence level $\alpha = 1 - 0.95 = 0.05$.
- In practice the interval is often constructed as a symmetric interval around a point estimate:
  - **point estimate±margin of error**
  - Rule of thumb: With a margin of error of 2 times the standard error you get a confidence interval, where the confidence level is approximately 95%.
  - I.e: **point estimate ± 2 x standard error** has confidence level of approximately 95%.
- Interpretation: We say that we are $100(1-\alpha)\%$ confident that the population parameter lies in the confidence interval.

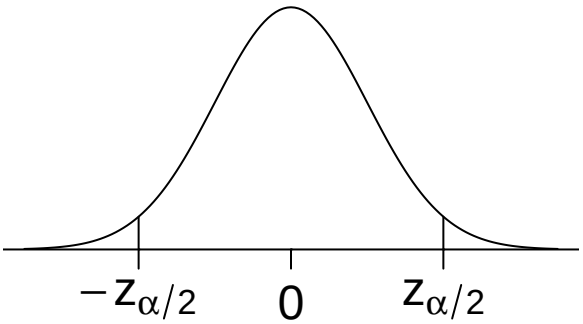
2.2 Confidence interval for the mean (known standard deviation)

- Consider a population with population mean μ and standard deviation σ. We would like to make a $100(1-\alpha)\%$ confidence interval for μ.
- Suppose we draw a random sample $X_1, \ldots, X_n$. As a point estimate for μ we use $\bar{X}$.
- If the population follows a normal distribution or if $n \geq 30$, we may assume $\bar{X} \sim \texttt{norm}(\mu, \frac{\sigma}{\sqrt{n}})$.
- The z-score of $\bar{X}$ follows a standard normal distribution:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \texttt{norm(0,1)}.$$

- We determine the **critical z-value** $z_{\alpha/2}$ such that $P(Z > z_{\alpha/2}) = \alpha/2$. This implies by symmetry that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$



- Inserting what Z is, we get

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

- Isolating μ in both in inequalities, we get

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

- That is, for $100(1-\alpha)\%$ of all samples, the population mean μ lies in the interval

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

2.3 Unknown standard deviation

- The confidence interval was derived using the z-score $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.
- Problem: In practice σ is typically unknown.
- If we replace σ by the sample standard deviation S, we get the **t-score**

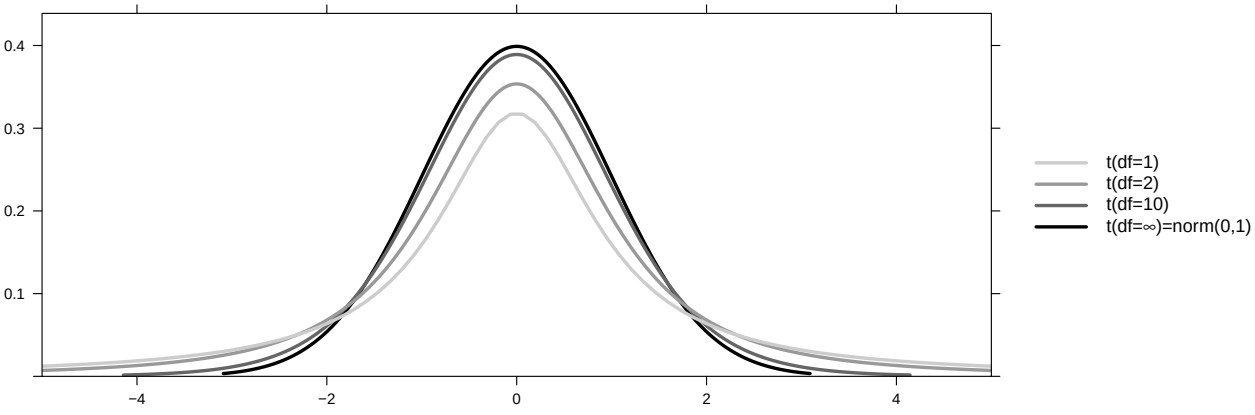
$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}}.$$

- Since S is random with a certain variance, this causes T to vary more than Z.
- As a consequence, T no longer follows a normal distribution, but a **t-distribution** with $n-1$ degrees of freedom.

2.4 t-distribution and t-score

- The **t-distribution** is very similar to the standard normal distribution:
  - it is symmetric around zero and bell shaped, but
  - it has "heavier" tails and thereby
  - a slightly larger standard deviation than the standard normal distribution.
  - Further, the t-distribution's standard deviation decays as a function of its **degrees of freedom**, which we denote df,
  - and when df grows, the t-distribution approaches the standard normal distribution.

The expression of the density function is of slightly complicated form and will not be stated here, instead the t-distribution is plotted below for $df = 1, 2, 10$ and ∞.



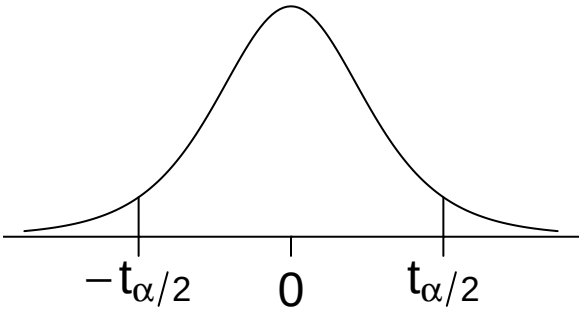
2.5 Confidence interval (unknown standard deviation)

- In the situation where σ is unknown, we use that

$$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim \mathtt{t}(n-1).$$

- We determine the **critical t-value** $t_{\alpha/2}$ such that $P(T > t_{\alpha/2}) = \alpha/2$. This implies by symmetry that

$$P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha.$$



- By exactly the same computations as before, we find that for $100(1-\alpha)\%$ of all samples, μ lies in

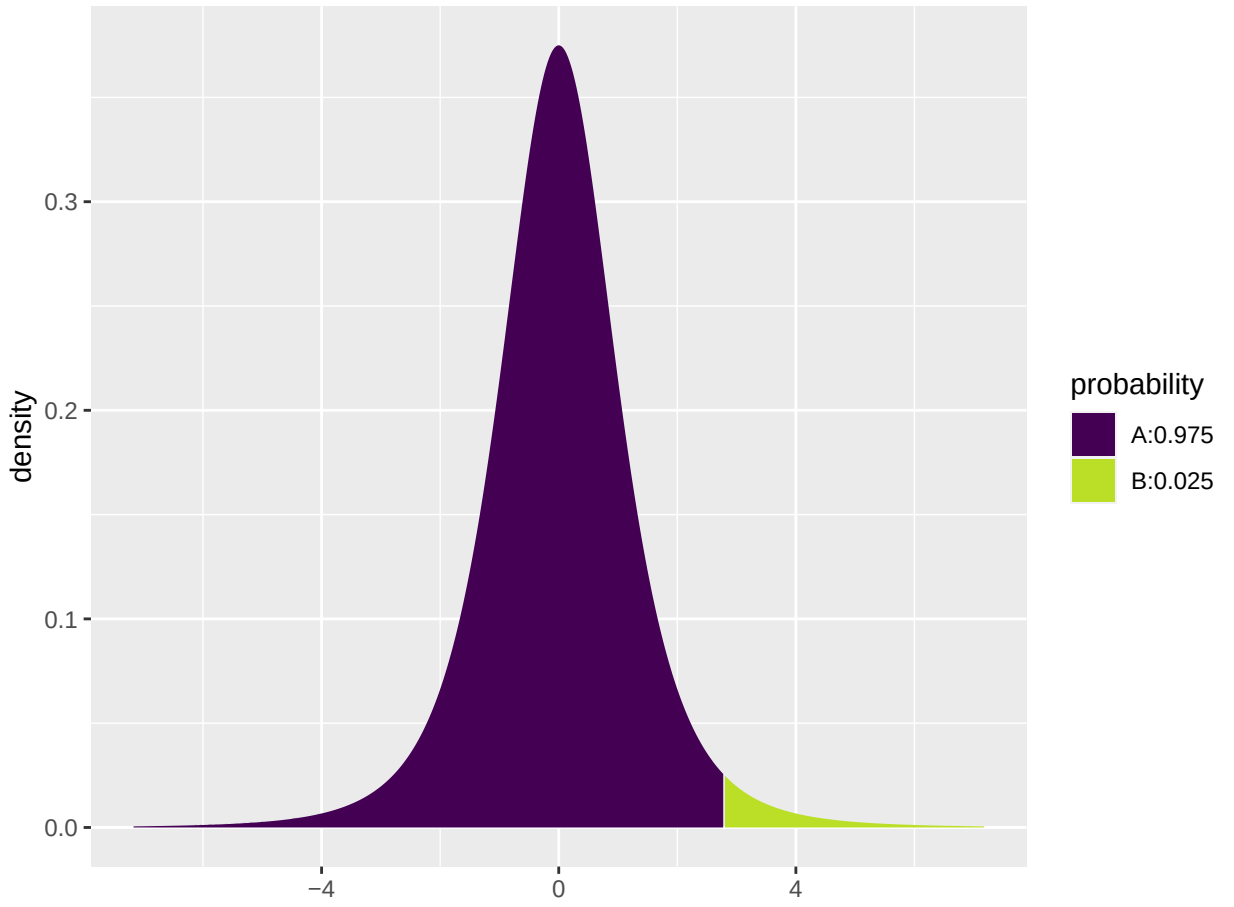
$$\left[\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right].$$

- This interval is what we call the $100(1-\alpha)\%$ **confidence interval** for μ.

2.6 Calculation of critical t-value in R

- To apply the formula, we need to be able to compute the critical t-value $t_{\alpha/2} = P(T > \alpha/2)$.
- This can be done in R via the function `qdist`.
- Note that we need the point with **right tail** probability $\alpha/2$ while R gives the **left tail** probailities. The right tail probability $\alpha/2$ corresponds to the left tail probability $1 - \alpha/2$.
- So to find $t_{\alpha/2}$ with $\alpha = 0.05$ (corresponding to a 95% confidence level) in a t-distribution with 4 degrees of freedom, we type:

```
qdist("t", p = 1 - 0.025, df = 4)
```




```
## [1] 2.776445
```

2.6.1 Example: Confidence interval for mean

- We return to the dataset `mtcars`. We want to construct a 95% confidence interval for the population mean μ of the fuel consumption.

```
stats <- favstats( ~ mpg, data = mtcars)
stats
```

```
##   min     Q1 median   Q3  max     mean       sd  n missing
##  10.4 15.425   19.2 22.8 33.9 20.09062 6.026948 32       0
```

```
qdist("t", 1 - 0.025, df = 32 - 1, plot = FALSE)
```

```
## [1] 2.039513
```

- I.e. we have
  - $\bar{x} = 20.1$
  - $s = 6$
  - $n = 32$
  - $df = n - 1 = 31$
  - $t_{crit} = 2.04$.
- The confidence interval is $\bar{x} \pm t_{crit} \frac{s}{\sqrt{n}} = [17.9, 22.3]$
- All these calculations can be done automatically by R:

```
t.test( ~ mpg, data = mtcars, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  mpg
## t = 18.857, df = 31, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  17.91768 22.26357
## sample estimates:
## mean of x
##  20.09062
```

2.6.2 Example: Plotting several confidence intervals in R

- We shall look at a built-in **R** dataset `chickwts`.
- `?chickwts` yields a page with the following information

  > An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens. Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Their weights in grams after six weeks are given along with feed types.

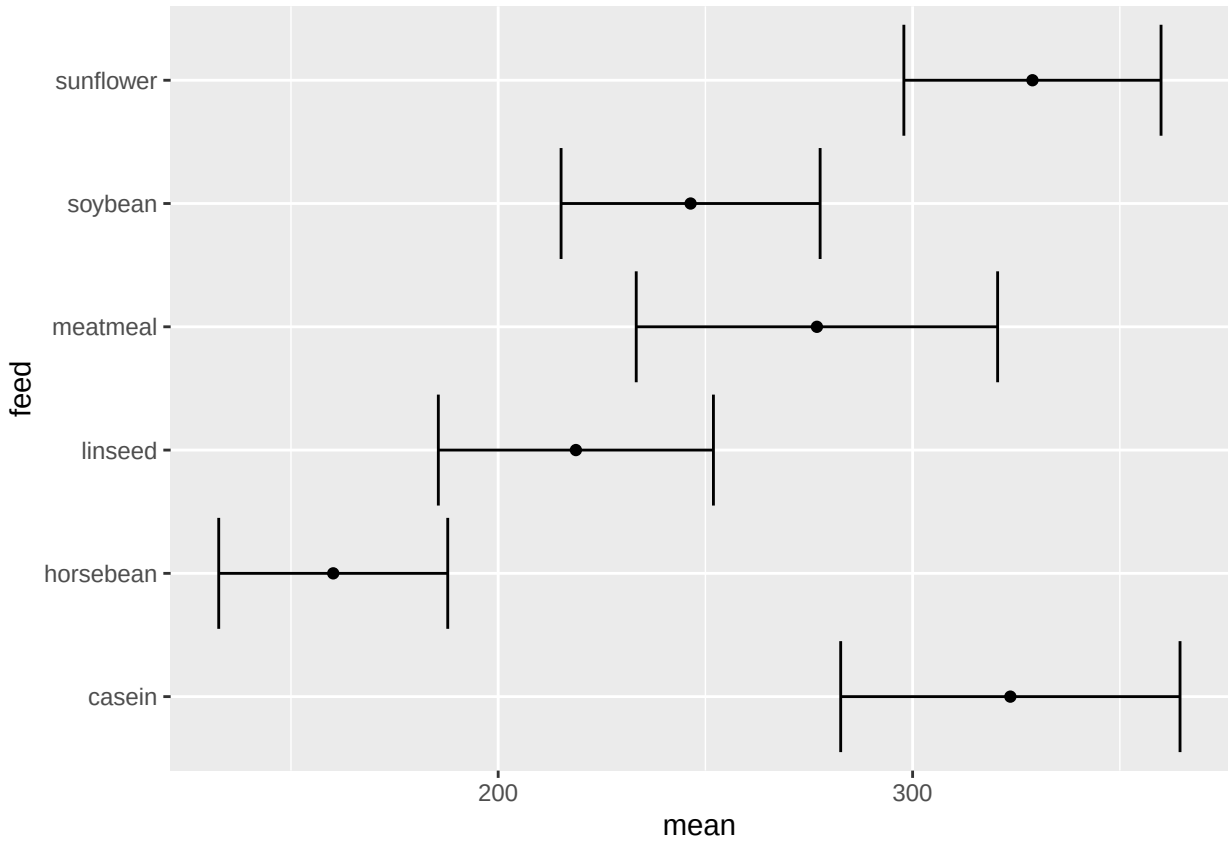
- `chickwts` is a data frame with 71 observations on 2 variables:
  - `weight`: a numeric variable giving the chick weight.
  - `feed`: a factor giving the feed type.
- Calculate a confidence interval for the mean weight for each feed separately; the confidence interval is from `lower` to `upper` given by `mean`$\pm$`tscore * se`:

```
cwei <- favstats( weight ~ feed, data = chickwts)
se <- cwei$sd / sqrt(cwei$n) # Standard errors
tscore <- qdist("t", p = .975, df = cwei$n - 1, plot = FALSE) # t-scores for 2.5% right tail probabilit
cwei$lower <- cwei$mean - tscore * se
cwei$upper <- cwei$mean + tscore * se
cwei[, c("feed", "mean", "lower", "upper")]
```

```
##        feed     mean    lower    upper
## 1    casein 323.5833 282.6440 364.5226
## 2 horsebean 160.2000 132.5687 187.8313
## 3   linseed 218.7500 185.5610 251.9390
## 4  meatmeal 276.9091 233.3083 320.5099
## 5   soybean 246.4286 215.1754 277.6818
## 6 sunflower 328.9167 297.8875 359.9458
```

- We can plot the confidence intervals as horizontal line segments using `gf_errorbarh`:

```
gf_errorbarh(feed ~ lower + upper, data = cwei) %>%
  gf_point(feed ~ mean)
```



2.7 Confidence interval for proportion

- Consider a population with a distribution that can only take the values 0 and 1. The interesting parameter of this distribution is the proportion p of the population that has the value 1.

- Given a random sample $X_1, \ldots, X_n$, recall that we estimate p by

$$\hat{P} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}.$$

- Based on the central limit theorem we have:

$$\hat{P} \approx N\left(p, \frac{\sigma}{\sqrt{n}}\right)$$

if both $n\hat{p}$ *and* $n(1-\hat{p})$ *are at least* 15.

- It can be shown that a random variable that takes the value 1 with probability p and 0 with probability $(1-p)$ has standard deviation

$$\sigma = \sqrt{p(1-p)}.$$

That is, the standard deviation is not a "free" parameter for a 0/1 variable as it is determined by the probability p.

- With a sample size of n, the standard error of $\hat{P}$ will be:

$$\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}.$$

- We do not know p but we insert the estimate $\hat{P}$ and get the **estimated standard error** of $\hat{P}$:

$$se = \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

- By similar calculations as in the case of confidence intervals for the mean, we find the limits of the confidence interval to be

$$\hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}.$$

- Here $z_{\alpha/2}$ (or z_{crit}) is the critical value for which the upper tail probability in the standard normal distribution is $\alpha/2$. (E.g. we have $z = 1.96$ when $\alpha = 5\%$.)

2.7.1 Example: Point and interval estimate for proportion

- We consider again the data set concerning votes in Chile.
- We are interested in the unknown proportion p of females in the population of Chile.
- The gender distribution in the sample is:

```
library(mosaic)
tally( ~ sex, data = Chile)
```

```
## sex
##    F    M
## 1379 1321
```

```
tally( ~ sex, data = Chile, format = "prop")
```

```
## sex
##         F         M
## 0.5107407 0.4892593
```

- Estimate of p (sample proportion): $\hat{p} = \frac{1379}{1379+1321} = 0.5107$
- An approximate 95% confidence interval for p is:

$$\hat{p} \pm z_{crit} \times se = 0.5107 \pm 1.96\sqrt{\frac{0.5107(1-0.5107)}{1379+1321}} = (0.49, 0.53)$$

- Interpretation: We are 95% confident that there is between 49% and 53% females in Chile.

2.7.2 Example: Confidence intervals for proportion in R

- R automatically calculates the confidence interval for the proportion of females when we do a so-called hypothesis test (we will get back to that later):

```
prop.test( ~ sex, data = Chile, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  Chile$sex  [with success = F]
## X-squared = 1.2459, df = 1, p-value = 0.2643
```

```
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4918835 0.5295675
## sample estimates:
##         p
## 0.5107407
```

- The argument `correct = FALSE` is needed to make R use the "normal" formulas as on the slides and in the book. When `correct = TRUE` (the default) a mathematical correction which you have not learned about is applied and slightly different results are obtained.

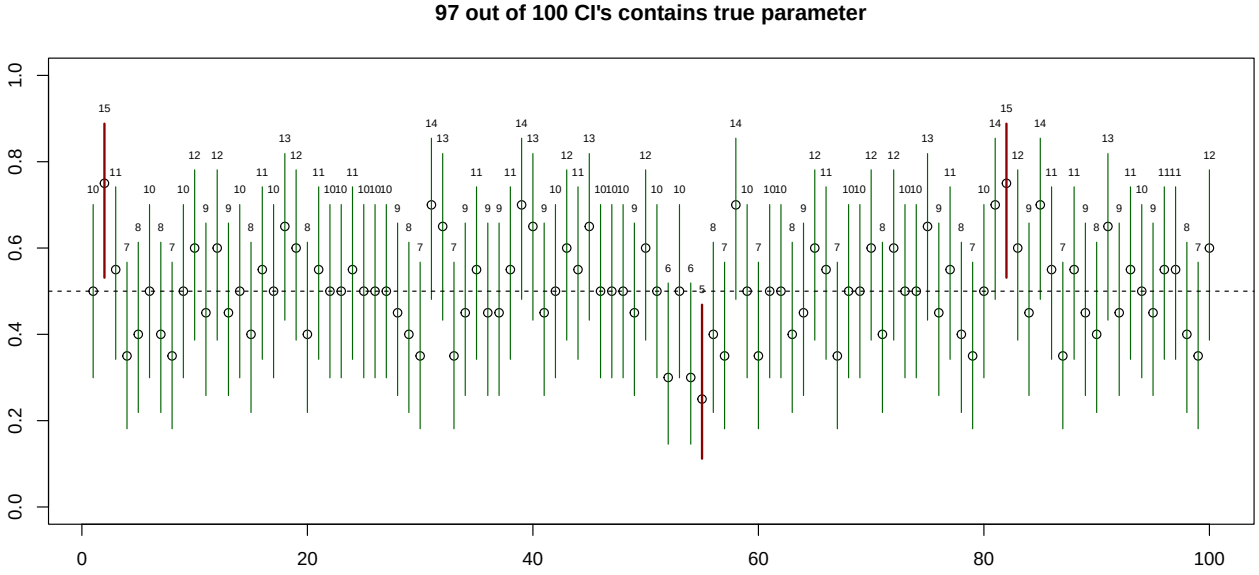
2.7.3 Example: Chile data

We could also have computed a 99% confidence interval for the proportion of females in Chile:

- For a 99%-confidence level we have $\alpha = 1\%$ and
  1) z_{crit}=`qdist("norm", 1 - 0.01/2)`=2.576.
  2) We still have $\hat{p} = 0.5107$ and $n = 2700$, so $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.0096$.
  3) Thereby, a 99%-confidence interval is: $\hat{p} \pm z_{crit} \times se = [0.49, 0.54]$.
- Note that the confidence interval becomes wider, when we want to be more confident that the interval contains the population parameter.

3 Confidence interval interpretation

- Assume a population parameter $p = 0.5$ (e.g. a fair coin)
- Draw 100 samples of size $n = 20$ and for each estimate $\hat{p}$ and 95% confidence interval



- Long run interpretation (repeating the experiment many, many times)
  - Each sample gives new CI limits
  - Before collecting data, the procedure for calculating a CI will provide a CI that with 95% probability will contain the true parameter value
  - Once data collected: Contains the true value or not
- CI contains values of the parameter that are compatible with the observed data
  - Either the CI contains the true parameter, or something "extraordinary" has happened (not impossible, just improbable/surprising)

- We are 95% confident that there is between 49% and 53% females in Chile

4 Confidence interval for the variance

4.1 The sample variance

- Suppose we are interested in the variance σ^2 of a population.
- We draw a random sample $X_1, \ldots, X_n$ and use the sample variance S^2 as a point estimate or σ^2.
- When the population distribution is normal, or $n \geq 30$,

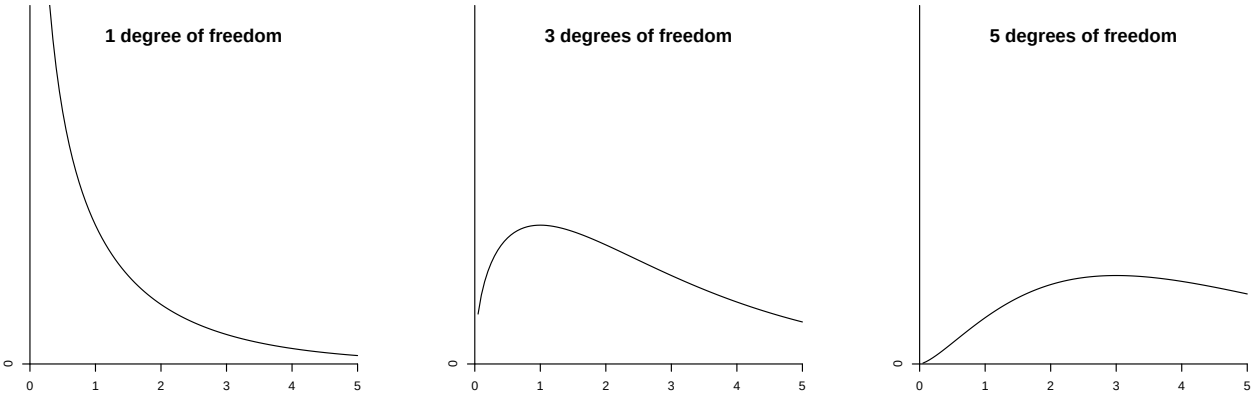
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

where $\chi^2(n-1)$ is the **chi-square distribution** with $(n-1)$ degrees of freedom (se next slide).

  - As a rule of thumb, the degrees of freedom are found as the number of observations (n) minus the number of unknown parameters describing the mean (one, namely μ).

4.2 The χ^2-distribution

- The distribution $\chi^2(k)$ is called the **chi-square** distribution.
  - It is a continuous distribution on $(0, \infty)$.
  - It depends on a the parameter k called the **degrees of freedom**.
  - The degrees of freedom determine the shape of the distribution.
  - The mean value is k.

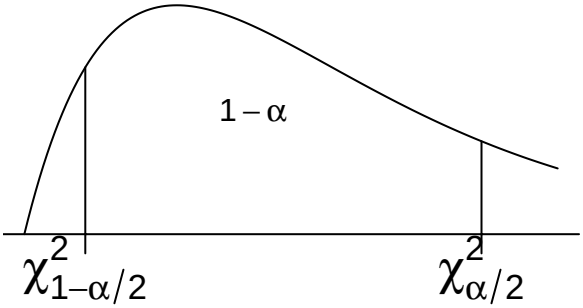


4.3 Confidence interval for variance

- To make a confidence interval for σ^2, we draw a random sample $X_1, \ldots, X_n$, compute S^2 and recall that

$$\frac{(n-1)S^2}{\sigma} \sim \chi^2(n-1).$$

- Let $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ be the critical values in a χ^2-distribution with $(n-1)$ degrees of freedom such the right tail probabilities are $\alpha/2$ and $1-\alpha/2$, respectively.



- Then

$$P\left(\chi^2_{1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}\right) = 1-\alpha.$$

- Isolating σ^2, this is equivalent to

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha.$$

- So we get the confidence interval for σ^2:

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}}; \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right].$$

- A confidence interval for σ can be found by taking square roots:

$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}}; \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}\right].$$

- Note that these confidence intervals are not symmetric around the point estimate.

4.3.1 Example: confidence interval for a variance

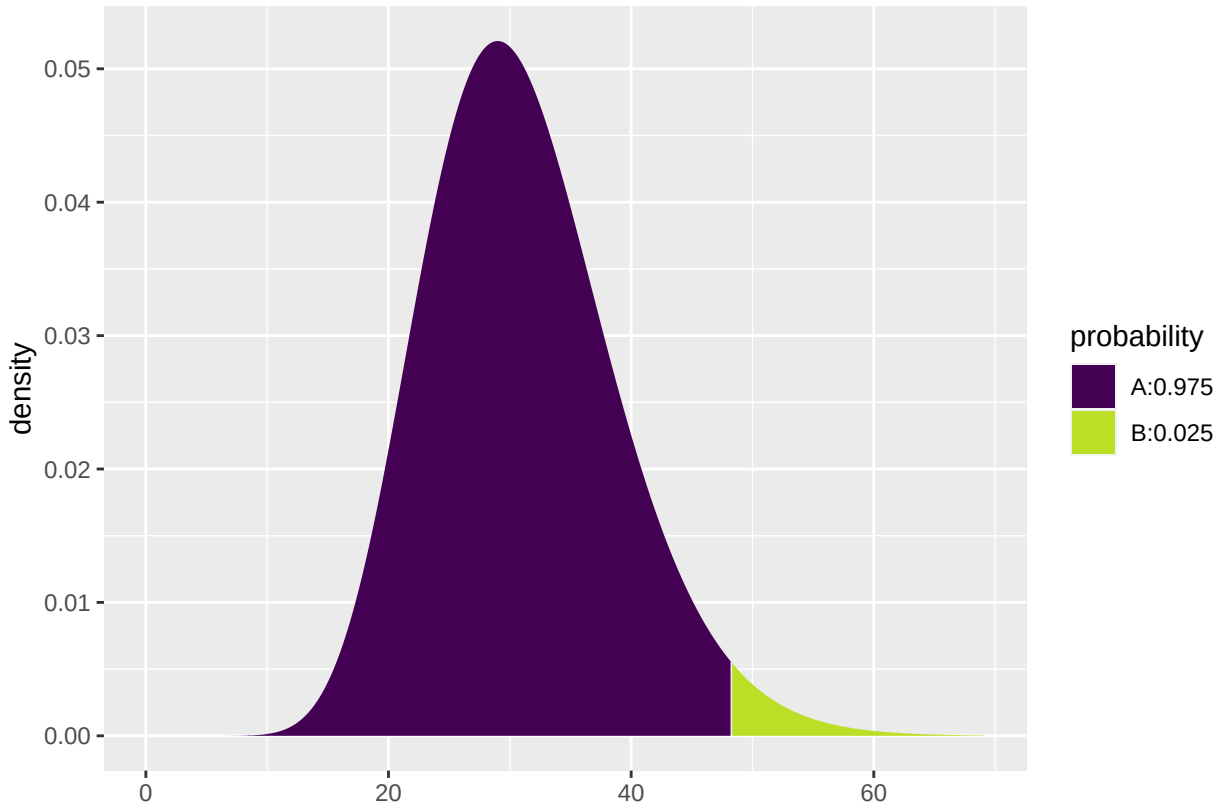
- We return to the dataset `mtcars` and construct a 95% confidence interval for the population variance σ^2 of the fuel consumption. We find the sample variance to be $6.026948^2 \approx 36.3$ using ´favstats´.

```
stats <- favstats( ~ mpg, data = mtcars)
stats
```

```
##   min     Q1 median   Q3  max     mean       sd  n missing
##  10.4 15.425   19.2 22.8 33.9 20.09062 6.026948 32       0
```

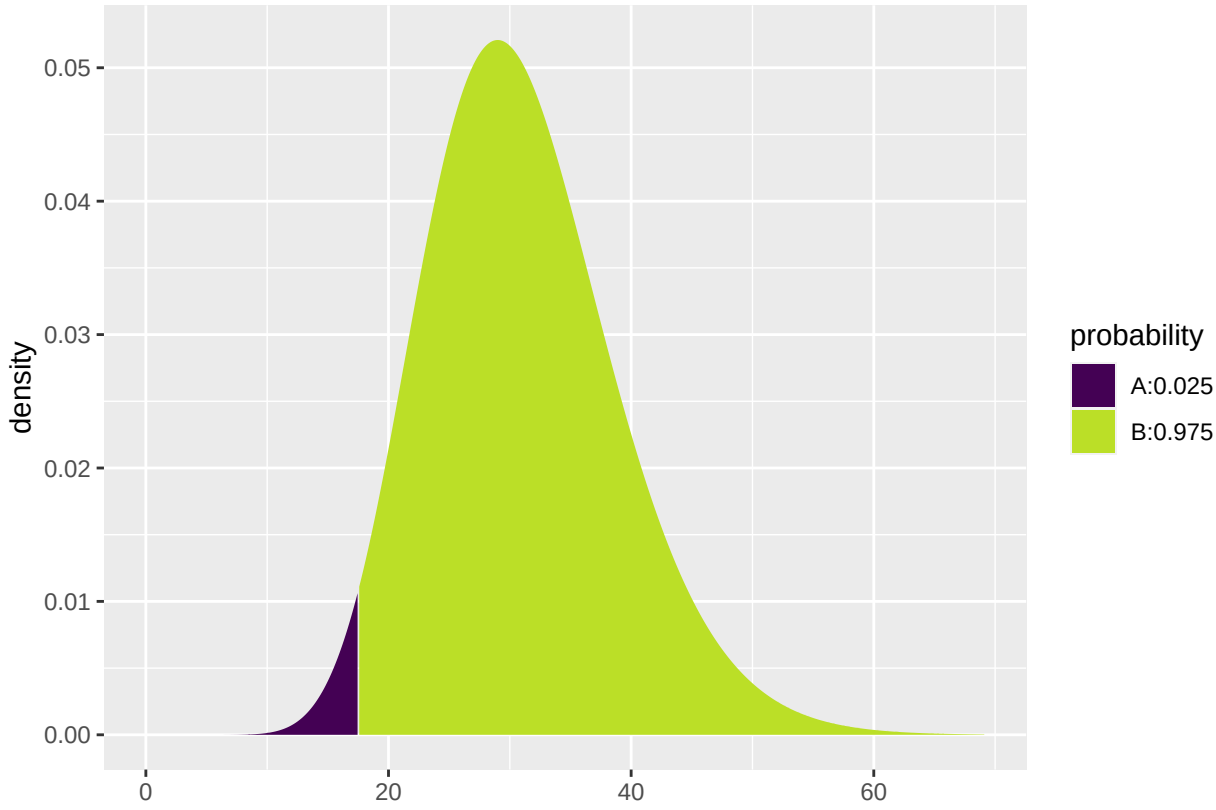
- The critical χ^2-values are found using `qdist`. The degrees of freedom are $n-1$. The χ^2-distribution is not symmetric, so we need to find both $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$.

```
qdist("chisq", 1 - 0.025, df = 32 - 1 )
```




```
## [1] 48.23189
```

```
qdist("chisq",  0.025, df = 32 -1)
```



```
## [1] 17.53874
```

- I.e. we have
  - $\chi^2_{\alpha/2} = 48.23189$
  - $\chi^2_{1-\alpha/2} = 17.53874$
- So we get the confidence interval for σ^2:

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right] = \left[\frac{31 \cdot 6.026948^2}{48.23189}; \frac{(31 \cdot 6.026948^2}{17.53874}\right] = [23.3, 64.2].$$

- A confidence interval for σ is $[\sqrt{23.3}, \sqrt{64.2}] = [4.83, 8.01]$.

5 Determining sample size

5.1 Determining sample size

- When planning an experiment, one has to decide how large the sample size should be.
  - If the sample size is too small, the parameter estimates will have high variance and hence confidence intervals will be large.
  - A sample size that is too large is costly in terms of time, money, etc.

5.2 Sample size for proportion

- The confidence interval is of the form point estimate±estimated margin of error.
- Imagine that we want to plan an experiment, where we **want to achieve a certain margin of error** M (and thus a specific width of the associated confidence interval).
- When we estimate a proportion the margin of error is

$$M = z_{crit}\sqrt{\frac{p(1-p)}{n}},$$

  where the critical z-score, z_{crit}, is determined by the specified confidence level.
- If we solve the equation above we see that if we choose sample size

$$n = p(1-p)\left(\frac{z_{crit}}{M}\right)^2,$$

  then we obtain an estimate of π with margin of error M.
- If we do not have a good guess for the value of p we can use the worst case value $p = 50\%$. The corresponding sample size $n = \left(\frac{z_{crit}}{2M}\right)^2$ ensures that we obtain an estimate with a margin of error, which is at the *most* M.

5.2.1 Example

- We want to make a survey to determine the proportion of the Danish population that will a certain party at the next election. How many voters should we ask to get a margin of error, which equals 1%?
- We set the confidence level to be 95%, which means that $z_{crit} = 1.96$.
- Worst case is $p = 0.5$, yielding:

$$n = p(1-p)\left(\frac{z_{crit}}{M}\right)^2 = \frac{1}{4}\left(\frac{1.96}{0.01}\right)^2 = 9604.$$

- If we are interested in the proportion of voters that vote for "socialdemokratiet'' a good guess is $p = 0.23$, yielding

$$n = p(1-p)\left(\frac{z_{crit}}{M}\right)^2 = 0.23(1-0.23)\left(\frac{1.96}{0.01}\right)^2 = 6804.$$

- If we instead are interested in "liberal alliance'' a good guess is $p = 0.05$, yielding

$$n = p(1-p)\left(\frac{z_{crit}}{M}\right)^2 = 0.05(1-0.05)\left(\frac{1.96}{0.01}\right)^2 = 1825.$$

5.3 Sample size for mean

- The confidence interval is of the form point estimate±estimated margin of error.
- Imagine that we want to plan an experiment, where we **want to achieve a certain margin of error** M.
- When we estimate a mean the margin of error is

$$M = z_{crit}\frac{\sigma}{\sqrt{n}},$$

  where the critical z-score, z_{crit}, is determined by the specified confidence level.
- If we solve the equation above we see:
  - If we choose sample size $n = \left(\frac{z_{crit}\sigma}{M}\right)^2$, then we obtain an estimate with margin of error M.
- Problem: We usually do not know σ. Possible solutions:
  - Based on similar studies conducted previously, we make a qualified guess at σ.
  - Based on a pilot study a value of σ is estimated.